NEXT GENERATION SEQUENCING ANALYSIS(ASSIGNMENT PART2)

Data Retrieval (NCBI SRA) & Introduction to data types

- Read Quality Check (FastQC & Cutadapt)
- Alignment of reads using reference Genome (BWA)
- Visualization of mapped reads (IGV)

Paired End-RAW DATA

Paired-end sequencing enables the sequencing of both ends of DNA fragments, producing high-quality, alignable sequence data. In this process, two files are generated: one containing all the forward sequences and the other containing the reverse sequences, with both files aligned in the same order. This method is particularly useful for detecting genomic rearrangements, repetitive sequence elements, gene fusions, and novel transcripts.

Treated – R1norm&R2norm

Control – R1TUMOR&R2TUMOR

The FASTQ format is a text-based method for storing biological sequences along with their corresponding quality scores. Each sequence in a FASTQ file spans four lines. The last line contains ASCII characters representing the quality scores of the bases in the second line. These quality scores, which identify nucleotide bases, are measured using the Phred quality score, indicating the probability of an error in base calling. The Phred scale's primary function is to automatically determine accurate, quality-based consensus sequences. For brevity, both the sequence letter and the quality score are encoded with ASCII characters.

FASTQC

FastQC offers an easy method for performing quality control checks on raw sequence data from high-throughput sequencing pipelines. It is compatible with data produced by Illumina platforms.

Installation of FASTQC in LINUX

```
sudo apt-get install fastqc
```

To get the Quality report of raw data

```
fastqc R1norm R2norm R1TUMOR R2TUMOR
```

If the read quality is satisfactory, you can proceed with the mapping process. If not, use Cutadapt software to clean the data.

CUTADAPT
Cutadapt identifies and eliminates adapter sequences, primers, poly-A tails, and other unwanted sequences from high-throughput sequencing reads.

To download and install Cutadapt

sudo pip install cutadapt

To trim a 3’ adapter, with AACCGGTT as the adapter to be removed, reads are processed from input.fastq and the results are saved in output.fastq.
cutadapt -a AACCGGTT -o cutout_1.fastq read1.fastq

To trim a 5’ adapter, with AACCGGTT as the adapter to be removed, reads are processed from input.fastq and the results are saved in output.fastq.
cutadapt -g AACCGGTT -o cutout_1.fastq read1.fastq

FOR paired end(R1norm&R2norm)

cutadapt -q 10 -a ADAPTER --minimum-length 20 -o tmp.1.fastq -p tmp.2.fastq R1norm.fastq & R2norm.fastq

cutadapt -q 15 -a ADAPTER --minimum-length 20 -o trimmed.2.fastq -p trimmed.1.fastq tmp.2.fastq tmp.1.fastq

FOR paired end(R1TUMOR&R2TUMOR)

cutadapt -q 10 -a ADAPTER --minimum-length 20 -o tmp.1.fastq -p tmp.2.fastq R1TUMOR.fastq & R2TUMOR.fastq

cutadapt -q 15 -a ADAPTER --minimum-length 20 -o trimmedtumor.2.fastq -p trimmedtumor.1.fastq tmptumor.2.fastq tmptumor.1.fastq

Again check the quality of reads; if everything is good, go ahead with mapping.

Mapping on Reference Genome using BWA.

DOWNLOAD THE REFRENCE GENOME FROM UCSC GENOME BROWSER. T2T-CHM13:
First, prepare the ‘hs1.fa.gz’ file.

To analyze sequence data, aligning it to a reference genome is essential. BWA is a software package designed for mapping low-divergence sequences against large reference genomes, like the human genome.
BWA first needs to construct index for the reference genome.

```
gunzip hs1.fa.gz
bwa index –a bwtsw hs1.fa
```

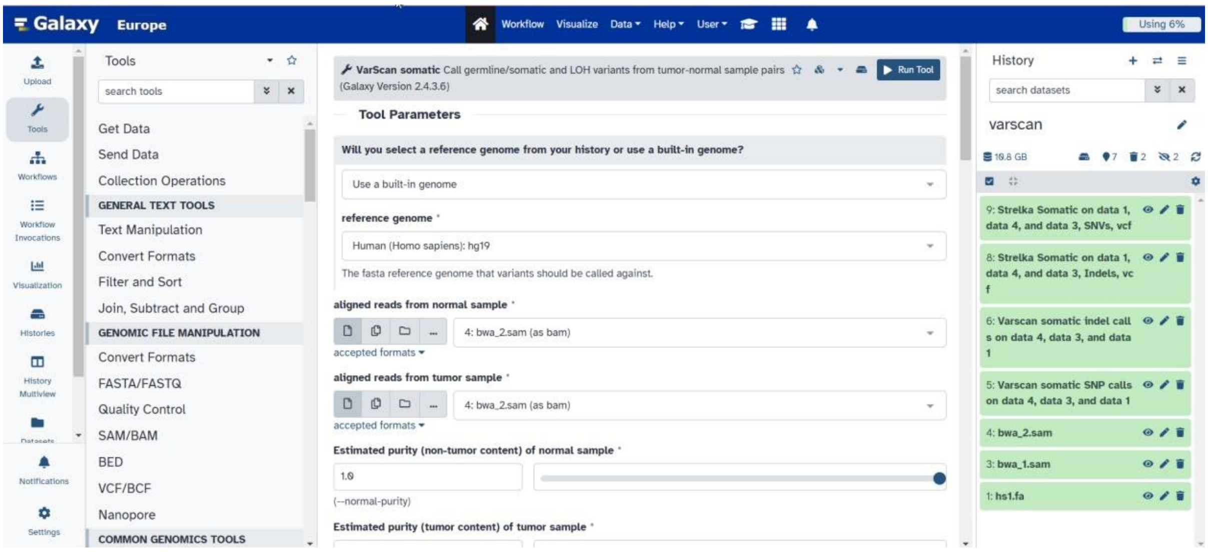
```
sudo fallocate -l 16G /swapfile
sudo chmod 600 /swapfile
sudo mkswap /swapfile
sudo swapon /swapfile
 free -m
```

Using Mem

To align the reads on the reference genome
For paired end:

```
bwa mem hs1.fa trimmed.1.fastq trimmed.2.fastq > bwa_1.sam
bwa mem hs1.fa trimmedtumor.1.fastq trimmedtumor.2.fastq > bwa_2.sam
```

The **GALAXY server** was utilized to run Strelka or VarScan for the purpose of identifying somatic mutations.
The tumour and normal SAM files, along with the reference genome file, were uploaded to the GALAXY server.



VarScan somatic Call germline/somatic and LOH variants from tumor-normal sample pairs
(Galaxy Version 2.4.3.6)

Tool Parameters

Will you select a reference genome from your history or use a built-in genome?
Use a built-in genome

reference genome *
Human (Homo sapiens): hg19
The fasta reference genome that variants should be called against.

aligned reads from normal sample *
4: bwa_2.sam (as bam)

aligned reads from tumor sample *
4: bwa_2.sam (as bam)

Estimated purity (non-tumor content) of normal sample *
1.0
(--normal-purity)

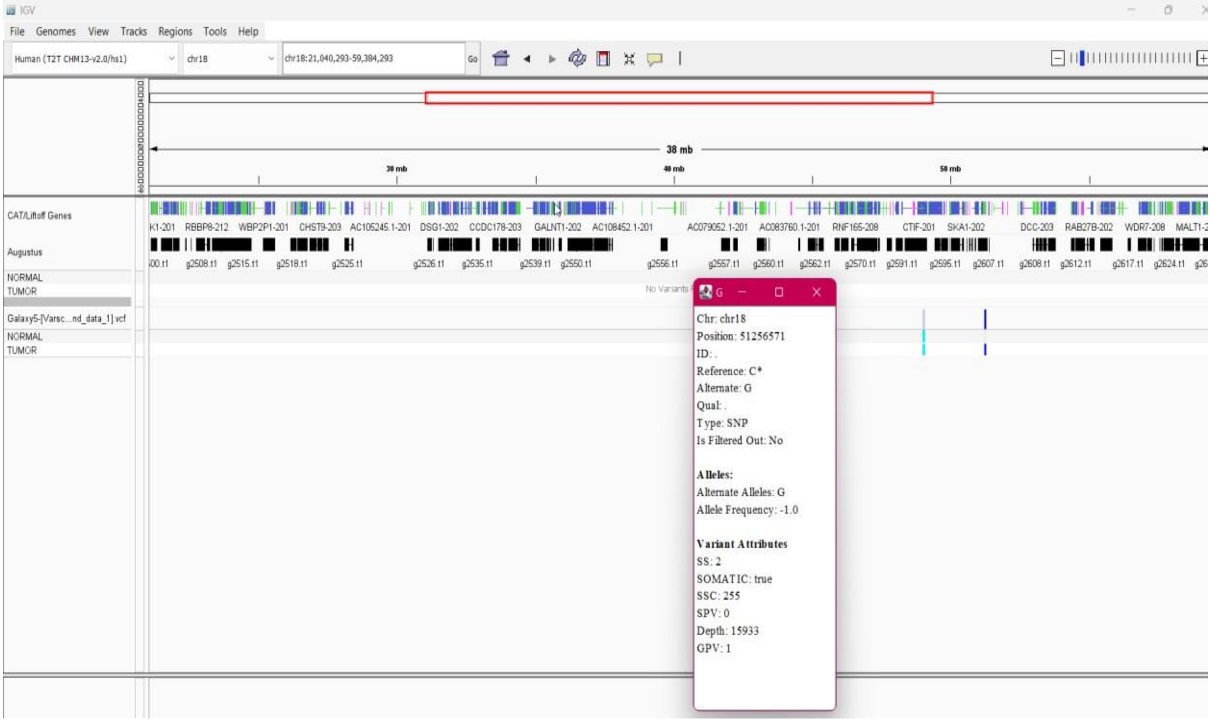
Estimated purity (tumor content) of tumor sample *

varscan
10.8 GB 7 2 2

9: Strelka Somatic on data 1, data 4, and data 3, SNVs, vcf
8: Strelka Somatic on data 1, data 4, and data 3, Indels, vcf
6: Varscan somatic indel calls on data 4, data 3, and data 1
5: Varscan somatic SNP calls on data 4, data 3, and data 1
4: bwa_2.sam
3: bwa_1.sam
1: hs1.fa

With four output files in VCF format, we can perform various analysis.
Galaxy9-[Strelka_Somatic_on_data_1,_data_4,_and_data_3,_SNVs,_vcf]
Galaxy8-[Strelka_Somatic_on_data_1,_data_4,_and_data_3,_Indels,_vcf]
Galaxy6-[Varscan_somatic_indel_calls_on_data_4,_data_3,_and_data_1]
Galaxy5-[Varscan_somatic_SNP_calls_on_data_4,_data_3,_and_data_1]

IGV (Integrative Genomics Viewer): A powerful visualization tool that allows you to view VCF files and explore genetic variations in a genome browser interface given below.


IGV
File Genomes View Tracks Regions Tools Help
Human (T2T CHM13-v2.0/hs1)
chr18
chr18:21,040,293-59,384,293
Go
38 mb
30 mb
40 mb
50 mb
CAT/Liftoff Genes
Augustus
NORMAL
TUMOR
Galaxy5-[Varsc...nd_data_1].vcf
NORMAL
TUMOR
No Variants
K1-201 RBBP8-212 WBP2P1-201 CHST9-203 AC105245.1-201 DSG1-202 CCDC178-203 GALNT1-202 AC108452.1-201 AC079052.1-201 AC083760.1-201 RNF165-208 CTIF-201 SKA1-202 DCC-203 RAB27B-202 WDR7-208 MALT1-2
00.t1 g2508.t1 g2515.t1 g2518.t1 g2525.t1 g2526.t1 g2535.t1 g2539.t1 g2550.t1 g2556.t1 g2557.t1 g2560.t1 g2562.t1 g2570.t1 g2591.t1 g2595.t1 g2607.t1 g2608.t1 g2612.t1 g2617.t1 g2624.t1 g26
G

Chr: chr18
Position: 51256571
ID: .
Reference: C*
Alternate: G
Qual: .
Type: SNP
Is Filtered Out: No

Alleles:
Alternate Alleles: G
Allele Frequency: -1.0

Variant Attributes
SS: 2
SOMATIC: true
SSC: 255
SPV: 0
Depth: 15933
GPV: 1